DOCUMENT RESUME

ED 212 128 FL 012 160

AUTHOR McLean, L. D.
TITLE Separating the Effect of More Hours of Instruction from General Language Development on Achievement in French as a Second Language.
INSTITUTION Ontario Inst. for Studies in Education, Toronto.
SPONS AGENCY Ontario Dept. of Education, Toronto.
PUB DATE [80]
NOTE 27p.; Revised version of a paper presented at the International Symposium on Educational Testing (4th, Antwerp, Belgium, June 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Communicative Competence (Languages); Elementary Education; Foreign Countries; *French; Language Proficiency; *Language Tests; Maturity (Individuals); Secondary Education; *Second Language Instruction; Speech Skills; Test Construction; Writing Skills
IDENTIFIERS Ontario (Toronto)

ABSTRACT

This paper presents research on the contribution of the amount of schooling to performance in French as a second language, assessed by newly developed measures of communicative competence. Test items addressed three components of communicative competence: grammatical, sociolinguistic, and strategic. Aware of the relationship of context to communication, test developers presented every task, or test item, within a meaningful context, and, whenever possible, language samples were taken from authentic materials. Test items, including both speaking and writing tasks, were prepared and administered to 6000 grade 6 and grade 10 students in 849 core French classes. Data are reported in terms of levels incorporating both maturity (growth from grade 6 to grade 10) and time (cumulative number of hours of instruction). The unit of analysis for performance evaluation is the classroom p-value (proportion correct). The analysis reveals that substantial gains are due to maturation rather than added instruction. (MES)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SUBMITTED TO THE CANADIAN JOURNAL OF EDUCATION

ED212128

Separating the Effect of More Hours of Instruction from General Language Development on Achievement in French as a Second Language

by

L. D. McLean
Department of Measurement, Evaluation & Computer Applications
The Ontario Institute for Studies in Education
Toronto, Ontario, Canada

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. D. McLean

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FL012140

Separating the Effect of More Hours of Instruction from General Language Development on Achievement in French as a Second Language

ABSTRACT

Over 6000 grade 6 or grade 10 students in 849 core French classes responded to 100 new French test items as part of the development of the Ontario Assessment Instrument Pool. Classes were found in which a majority of students had accumulated either 180, 360, 540 or 720 hours of instruction in French. Most items were administered in all classes. Using the classroom p-value as the unit of analysis, gains due to more instruction, regardless of grade, were separated from gains observed from grade 6 to grade 10 for the same amount of instruction. Limitations on the study require cautious interpretation of findings, but the analysis did reveal substantial gains due to maturation (holding amount of instruction constant) and very modest gains with added instruction (holding maturation constant). It was suggested that the myriad experiences that contribute to first language development also contribute to improved second language performance.

Separating the Effect of More Hours of Instruction from General Language Development on Achievement in French as a Second Language

Language teaching and language testing are today exciting areas of research, development and practice. This is so for language as a mother tongue but especially so for "second" or "foreign" languages. Perhaps there has always been such excitement in multilingual countries, but in North America and the UK at least, our emotions have only recently been aroused.

It is especially exciting for some of us that developments in language testing are keeping pace with those in language theory. Indeed, a theory that cannot be tested is of dubious worth, so theories of communicative competence must also have tests of communicative competence if they are to be useful.

The work on which this paper is based actually evolved in the reverse order--a request for tests stimulated production of a refined theory (Canale and Swain, 1980) which was then used to build some new measuring instruments. The project of which this work is a part has been described elsewhere (Research and Evaluation Branch, 1979), so only two highlights of the theory will be mentioned, both essential to this presentation.

First, three components of communicative competence were defined: Grammatical, Sociolinguistic and Strategic. Grammatical competence is more traditional, "knowledge of lexical items and rules of morphology, syntax, sentence-grammar semantics, and phonology" (Canale and Swain, 1980, p. 29). Sociolinguistic competence includes the ability to say the appropriate thing at the appropriate time, to "get the message

across," and is a recognition that grammatical correctness is rarely sufficient, or even necessary, to communication. Knowing and using alternative ways to communicate when you don't know exactly what to say or write is the essence of the third component, strategic competence. This framework has profound implications for teaching and it has met the objective of giving shape and direction to the development of measuring instruments.

The theory also stresses the fundamental importance of context to communication--not an entirely new insight, but one that has been somewhat neglected in teaching and almost entirely neglected in testing. It was decided that every task, every test item if you will, would be presented within a meaningful context and that wherever possible, language samples would be taken from language-in-use--newspapers, radio programs and the like. A survey of existing Canadian tests revealed that for the most part they employed artificial contexts or supplied no context at all. A group in the United Kingdom recently condemned tests for the same reason (Moys, Harding, Page and Printon, 1980).

The focus of the present paper is on the contribution of schooling, especially the amount of schooling, to performance on newly developed measures of communicative competence. Before describing the empirical study, we review some recent research that helps motivate attention to amount of schooling and suggests other needed refinements.

GRACENOTES IN THE MUSIC OF TIME

Total time spent learning is being seen more and more as an important indicator of success in school subjects, and nowhere more than

in learning a second language. The school situation in second language is complicated because many students have the option to take more intensive second-language (more time per day or per week) and because students do move from school to school. Language teachers must learn to cope with groups of students having widely differing number of hours of instruction. Individual classes may be relatively homogeneous but vary greatly from place to place and year to year.

In the province of Ontario, for example, students in English schools begin the study of French anywhere from Kindergarten (age 4 or 5) to grade 7 (age 13 or 14), in programs that range from "immersion" (schooling predominantly in French) to twenty minutes per day. Thus, a grade 6 classroom in one jurisdiction might contain students competent enough to follow their studies in French while grade 6 classes in another would enrol students whose prior schooling in French amounts to one year at 20 minutes per day. There are classes at most points between these extremes.

In 1979, the Ontario Ministry of Education introduced the requirement that a record be kept for each student of the student's cumulative number of hours of instruction in French, to be updated annually. An elementary school year at 20 minutes per day is counted as 60 hours, hence 3 years of such a program (grades 4, 5 and 6, for example) yields a cumulative total of 180 hours. (The Ontario school year is 10 months at the elementary level.)

Cumulative hours of instruction has come to play a role in the selection of textbooks, by the inclusion of prerequisites in the list of

approved texts. Moreover, school boards can qualify for additional grants (financing) by increasing the number of hours of French they offer. A provincial commission gave impetus to the use of cumulative time by suggesting 1200 hours as the minimum required for English speakers to achieve a basic level for everyday purposes (Gillin Commission, 1974). In Ontario at least, time spent learning is not just of academic interest.

Research interest has shifted to the distribution of teaching time, however, since time alone has proved to be rather too blunt a predictor. Primary French in the Balance (Burstall et al., 1974) is a particularly widely known report of a British study that concerned itself with the best age to begin instruction, comparing the performance of younger pupils with that of more mature pupils, keeping instructional time and intensity constant. Intensity was low--20-30 minutes per day. The authors concluded that time spent learning French after age 11 was more productive than the time spent learning earlier in elementary school. A more recent Ontario study (Morrison, Walsh, Pawley and Bonyun, 1979b) compared "early immersion" (100% French in K-1, 80% 2-5, 50% 6-12) with a 30-minute per day program through grade 5 or 6 followed by "late immersion" (80-100% in grades 6-7 or 7-8). In contrast to Primary French in the Balance, the early immersion students performed better than the late immersion students on every measure, though both groups performed well. Looking only at these two studies one would conclude that intense study is more valuable early while 20-30 minutes per day is more valuable later. Such conclusion would be contradicted by results from Quebec (Genesse, 1979) that late immersion students performed better.

Other researchers varied total time while controlling for maturity, studying only grade 7 students in the program of 20 to 40 minutes per day referred to in Ontario as "core French" (Morrison, Walsh, Pawley and Bonyun, 1979a). On both listening and speaking tests, students who had cumulative total of about 930 hours by grade 7 outperformed those who had only about 585. The degree of intensity of the students' instruction was unclear, since a pupil who steadily accumulates 930 hours by grade 7 has to have experienced a minimum intensity of more than 40 minutes per day from grade one.

Though the foregoing suggests that more hours of schooling result in better achievement, other studies suggest that a period of late intensity is able to overcome a substantial deficit in cumulative total. A yet unpublished survey (Lapkin, Swain, Kamin and Hanna, 1980) found that on a listening test and a cloze test, a late immersion group in grade 8 (N=87) that had accumulated only about 700 total hours of instruction (including the immersion) outperformed another grade 8 group (N=123) who had steadily accumulated a total of 1620 hours. Attributing the better performance solely to total hours of instruction would be unwise under the best of conditions, of course, and in this study there were complicating factors. The better performers (late immersion group with fewer total hours) had a mean IQ of 118, whereas nothing is known about the IQ of the comparison group. From a previously cited study (Morrison, 1979a), we know that among similar students the language tests correlated about .35 with IQ, so a lower IQ in the comparison group would necessitate an adjustment to the difference between groups.

The listening test differences was 2.42 ($t = 6.6$, $p < .01$), however, so only a large downward adjustment would change the result. Given the large difference in cumulative hours in the opposite direction, the result is of interest whatever the IQ findings.

The implication of these results (and many others, of course) is that we must concern ourselves not just with teaching time but with distribution of teaching time, as Professor H. H. Stern emphasized in his Twentyman Lecture to the British Modern Language Society (Stern, 1978).

THE FSL PROJECT--ANOTHER CHANCE FOR A TEST OF TIME

In Ontario, as was noted above, English-speaking students can elect a variety of French programs--among them early immersion, late immersion and partial immersion. The vast majority of students, however, are in core French--20-40 minutes per day starting in grade 4 or grade 5. In many countries these are called "foreign language" classes.

The Ontario Ministry of Education issues a "Guideline" for each subject at various ages by way of shaping the content and objectives of the curriculum, and schools are obliged to follow the guideline. In preparation for a new FSL quideline (recently issued) the Ministry contracted with O.I.S.E. to prepare a set of assessment instruments for core FSL, the first comprehensive set to be prepared centrally since provincial examinations were abolished in 1970. A beginning was made by concentrating on grade 6 and grade 10, covering reading, writing, speaking and listening.

About one hundred separate instruments were prepared and submitted to students for the first time in November and December, 1979. Analysis of data from this screening trial is reported below in terms of "levels" that incorporate both "maturity" (growth from grade 6 to grade 10) and "time" (cumulative number of hours of instruction). We know too little about the intensity of students' prior instruction to bring it into the analysis, but in the middle levels at least we can regard the instruction as having been mainly of low intensity.

There are many other variables not measured, motivation, actual time on task and the like, so all of these remain as competing candidates to explain any results. The number and diversity of schools involved (see below, 'Screening Trials . . .'), however, makes it less plausible that differences between classes having differing hours of instruction could be attributed solely to motivation or aspects of instruction. What we have is a non-experimental cross-sectional analysis of differences and change over time, a type of analysis with many weaknesses and few strengths. There is so little data of this kind, however, that these tentative results are of potential interest should subsequent research confirm and strengthen their validity.

Definition of Levels

It was decided not to attempt to write instruments for students having less than 180 hours of instruction (e.g., less than 3 years at 20 min. per day), so the following levels were defined:

Level 1: 180 hours by the end of grade 6 (approx. age 12)

Level 2: 360 hours by the end of grade 6

Level 3: 540 hours by the end of grade 6

Level 4: 360 hours by the end of grade 9 (approx. age 15)

Level 5: 540 hours by the end of grade 9

Level 6: 720 hours by the end of grade 9

Obviously, these can apply only approximately to a given student at a point in time, but it was hoped they would be useful benchmarks for instrument development and would prove helpful to teachers in selecting testing instruments for their particular classes.

Screening Trials--The Sample of Classes

Most instruments were tested at all levels, since no experience was available to suggest strongly otherwise. (Instruments to be discussed here are essentially long test items, so the more common term, item, will be used below). Schools in eight different school districts were contacted where records indicated that grade 6 or grade 10 classes had the required number of hours of instruction. Batteries of items were then assembled and sent out to classes so that, except for speaking items, about 100 students would respond to each item at each level. No random sampling processes were employed, so that these results are not generalizable either to the province or to the levels. The purpose was primarily item development, so districts were selected to reflect certain salient realities (the cities of Sudbury and Ottawa were included, for example, because they have large French-speaking populations). It deserves repeating

that these results require corroboration before application--preferably corroboration by means of a full-scale field test with a probability sample of students.

Screening Trials--Measures and Unit of Analysis

There being no "tests" to score, responses to each item were simply collected and coded by level, district (school board) and classroom. Classroom p-values (proportion correct) were calculated for multi-choice items and classroom means for speaking and writing items. Thus, the classroom is the unit of analysis for each item. For the items used in this analysis, the number of classrooms was 847.

The classroom p-value is an attractive measure since proportion of students answering correctly has a direct implication for learning and the classroom is the unit of instruction. An increase of 10 "points" in p-value (from .50 to .60, for example) translates into 10% more students getting the item right, a small but worthy achievement. Ratings of speaking and writing items have no such natural metric, but each item in the FSL pool is keyed to the FSL curriculum domain so that teachers can interpret performance on an item in terms of skills. For this analysis, the data on speaking and writing were examined for increases in ratings with increases in level.

RESULTS

The number and range of classroom p-values can be seen in figure 1, where results are pooled from 33 multiple-choice items and presented by level. Figure 2 presents results from seven of the 33 selected

because they showed a clearer pattern of growth over levels. Three of the seven were grammatical items and four sociolinguistic.

A two-way ANOVA on p-values (level by competence) showed level, competence and level x competence interaction all to be significant in the data from figure 2 (Table 1). The "level" effect was virtually guaranteed, of course, since items were selected which had shown growth. Several a priori contrasts among levels are implied by the concepts of maturity and time, and these proved informative in interpreting the significant interaction between level and competence. Because p-values are proportions, they are not likely to be normally distributed and hence may violate an ANOVA assumption. The same analysis was done on logarithms of p-values (more likely to meet assumptions), with identical results.

Definition and Meaning of Contrasts among Levels

The effect of "maturity" can be defined as a contrast, a difference, between average performance in Grade 10 and average performance in Grade 6, controlling for number of hours of instruction:

Maturity = (level 4 + level 5) - (level 2 + level 3).

One can verify from the definition of levels that this is a difference between older and younger groups having the same aggregate amount of instruction (360 + 540).

Similarly, one can define two contrasts that reflect time differences, controlling for grade. The effect of cumulative

instructional time, "Hours," controlling for maturity, is a difference between hours within the same grade.

Hours (180) = (level 3 - level 2) + (level 5 - level 4)

Hours (360) = (level 3 - level 1) + (level 6 - level 4)

Only Hours (180) is shown in Table 1, the result being the same if Hours (360) was used with more levels.

Size of Contrasts and Significance Tests

By examination of the ANOVA results at the level of the contrasts, we can see that the interaction of level and competence is a function of maturity but not of hours (maturity x competence interaction: $F = 5.14$, $p = .020$). Table 2 contains average p-values by competence and grade, suggesting substantial growth from grade 6 to grade 10 in grammatical competence (as revealed by higher proportions of successes) but only modest growth in sociolinguistic competence over that period. Number of hours of instruction was held constant.

Both Hour (180) and Hour (360) contrasts were not statistically significant, that is, the null hypothesis of zero difference could not be rejected, and neither showed a significant interaction with competence. The observed differences are also shown in Table 2.

Speaking Task Rating Categories and Scoring

Four scales were defined for all speaking task responses, with a fifth applied where a long response was required. Too few items were scored on the fifth scale to warrant analysis here. The four main scales were:

1. Information--A score is assigned representing the extent to which the student included the relevant information in the response, regardless of the adequacy of its expression.
2. Grammaticality--This scale reflects an evaluation of the grammatical or structural accuracy of the response. Points are subtracted from a maximum score (minimum zero) for a range of major grammatical errors.
3. Pronunciation--For repetition items, detailed phonetic boundaries were identified and a point assigned for successful pronunciation of each. For other speaking items, an overall score for pronunciation and intonation was assigned.
4. Appropriateness--This reflects two aspects of appropriateness of response--the respect paid to the "tu/vous" distinction and the amount of polite or appropriate elaboration provided.

Three members of project staff scored a subset of the items independently, discussed the disparities and refined the procedures. The same group then did all the scoring. No systematic estimation of rater reliability was attempted in this early stage of development, but the small group working together over many months were able to achieve near consensus.

Speaking Task Results

Because speaking tasks were administered individually, it could be arranged that many students did the same set of tasks (N=78 or 96).

This permitted estimation of correlations among the scales and hence a check on the degree of overlap in ratings. With few exceptions, the correlation matrices were characterized by small, positive, non-significant correlations ($r \leq 20$). Of 37 tasks (items), only two showed a pattern of significant correlations. Thus, the rating scales appear to reflect sufficiently different characteristics to justify examining them separately. Low reliabilities would also produce low correlations and this possibility cannot be excluded.

Figure 3 is different from figures 1 and 2, in that it presents gains rather than classroom means or medians. The gains are expressed in the same sort of contrasts as Maturity and Hours, this time as contrasts among classroom mean ratings.

The pattern on the left in figure 3 (maturity) suggests that there was substantially higher performance in grade 10 in ability to convey information, compared to grade six students with the same number of hours of instruction. The box on the right in figure 3 [Hours (180)] suggests that a much more modest gain, if any at all, can be attributed to 180 additional hours of instructions. The same can be said with less emphasis about grammaticality, but the pronunciation and appropriateness scores are not higher in grade 10 than in grade 6.

DISCUSSION

Results from the maturity vs. hours analysis are consistent with those in Primary French in the Balance (Burstall, et al., 1974). Given the same amount of instruction in French, older English-speaking students do better than younger ones. Unexpected, however, was the size of the difference attributable to maturity compared to that attributable to an extra 180 or 360 hours of instruction. Recall that for many students (those who have 20 min./day), 180 hours is three years of core French. If these results could be replicated, they would suggest something of the order of 10% more students might get test items correct if the number of hours of instruction were increased by 180, or even 360 hours. If the same amount of instruction is given to older students, then something like 40% more students will succeed on items such as these.

Remembering all the qualifications already cited, a plausible interpretation is that a large part of the improved performance that is evident in grade 10 over grade 6 is due to general language development. An additional 20 minutes per day of instruction over several years results in few, if any, performance gains. (Recall that results are quite different for periods of immersion, either early or late.) It may be, for example, that "extended" French is not yet intense enough to yield more than very modest gains. Drawing such conclusions from any one study would be unwise, but it does suggest some needed research.

Superior results in grammatical competence over socio-linguistic are primarily interpreted as evidence for the validity of the competence categories, since it is known that present teaching

methods favour grammatical skills. It appears that the instrument development team did succeed in operationalizing the theory--at least so far as these seven items were concerned. Finding more improvement where teaching is concentrated suggests that "maturity" does not mean students improve just by getting older. It may mean that there is even more concentration on grammatical competence in the upper grades. On the other hand, grammatical competence may be a trait showing more growth over time than sociolinguistic competence.

Results from the speaking and writing items were quite consistent with those from reading and listening, not an inevitable outcome given the very different nature of the tasks and the different scale of measurement. That students showed most improvement in giving information is seen as more support for the validity of the scoring categories, since more mature students are expected to have more information in general and to be better able to communicate it. That grammatical competence improvement in reading and listening do not seem to be matched by comparable improvement in the grammaticality of speech is an intriguing, if speculative finding. The different scales of measurement preclude direct comparisons of quantity.

We began with a sense of excitement. It may be somewhat tempered by the results, but then that so often happens when expectations are tested against data. There is much to remain excited about, though, in the many opportunities the theory suggests for communicative teaching.

NOTE

The work on which this paper is based was funded by the Ontario Ministry of Education, Research and Evaluation Branch, as part of the development of the Ontario Assessment Instrument Pool, and by the Ontario Institute for Studies in Education (OISE). The views expressed therein are the author's, however, and do not necessarily reflect those of the Ministry of Education or the OISE. This is a much revised version of a paper presented at the Fourth International Symposium on Educational Testing, Antwerp, Belgium, June, 1980. Sharon Lapkin, Jim Cummins and Michael Canale provided helpful comments on an earlier draft.

REFERENCES

Burstall, C., Jamieson, M., Cohen, S., & Hargreaves, M. Primary French in the Balance. Windsor, Berks: NFER Publishing Company Ltd., 1974.

Canale, M., & Swain, M. The theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1980, 1(1), 1-47.

Genesse, F. A comparison of early and late immersion programs. McGill University, 1979 (mimeo).

Gillin Commission. Report of the Ministerial Commission on the Teaching of French. Toronto: Ontario Ministry of Education, 1974.

Lapkin, S., Swain, M., Kamin, J., & Hanna, G. Report on the 1979 evaluation of the Peel County late French immersion program, grades 8, 10, 11 and 12. OISE, May, 1980 (mimeo).

Morrison, F., Walsh, M., Pawley, C., & Bonyun, R. French proficiency of grade 7 core students in Ottawa and Carleton board schools, Spring 1979. French Working Paper No. 124, Research Centre, Ottawa Board of Education, July, 1979a.

Morrison, F., Walsh, M., Pawley, C., & Bonyun, R. French proficiency of grade 8 primary and late-entry immersion students, Ottawa and Carleton boards, Spring 1979. French Working Paper No. 131, Research Centre, Ottawa Board of Education, December, 1979b.

Moys, A., Harding, A., Page, B., & Printon, V. Modern Language Examinations at 16-plus: A critical analysis. London: CILT, 1980.

Research and Evaluation Branch. The Ontario Assessment Instrument Pool--A Curriculum-based Aid to Evaluation. (Review and Evaluation Bulletins, Vol. 1, No. 1). Toronto: Ontario Ministry of Education, 1979. (Disponible en français: La Banque d'instruments de mesure de l'Ontario.)

Stern, H. H. Of honeymoons, hangovers and bootstraps in foreign language teaching: The Twentyman Lecture. Royal Festival Hall, London, December 16, 1978. OISE, 1978 (mimeo).

Table 1

Analysis of Variance Table with Classroom p-value as Dependent Variable and Two Factors, Level and Type of Communicative Competence (exact Least-squares Solution Using the MANOVA Program in PSTAT). Levels 2-5 are Included and only Grammatical and Socio-Linguistic Competence

Sources		df	F	p
Level		3		
Maturity	1		50.16	.001
Hours (180)	1		.04	ns
Mat. x Hours	1		.07	ns
Competence		1	8.65	.01
Level by Competence		3		
Mat. x Competence	1		5.14	.02
Hours x Competence	1		.24	ns
Mat. x Hours x Comp.	1		.15	ns
Residual		113		

Table 2

Average p-values and Estimates of Contrasts for the Interpretation of Significant Effects in ANOVA: Maturity x Competence and Time

Mat. x Competence:	Grammatical	Socio-linguistic
Grade 6 (L2 + L3)	.40	.38
Grade 10 (L4 + L5)	.75	.57

Time:	Grade 6	Grade 10
Hours (180)*	.00	.16
Hours (360)*	.08	.12

std. error 0.05

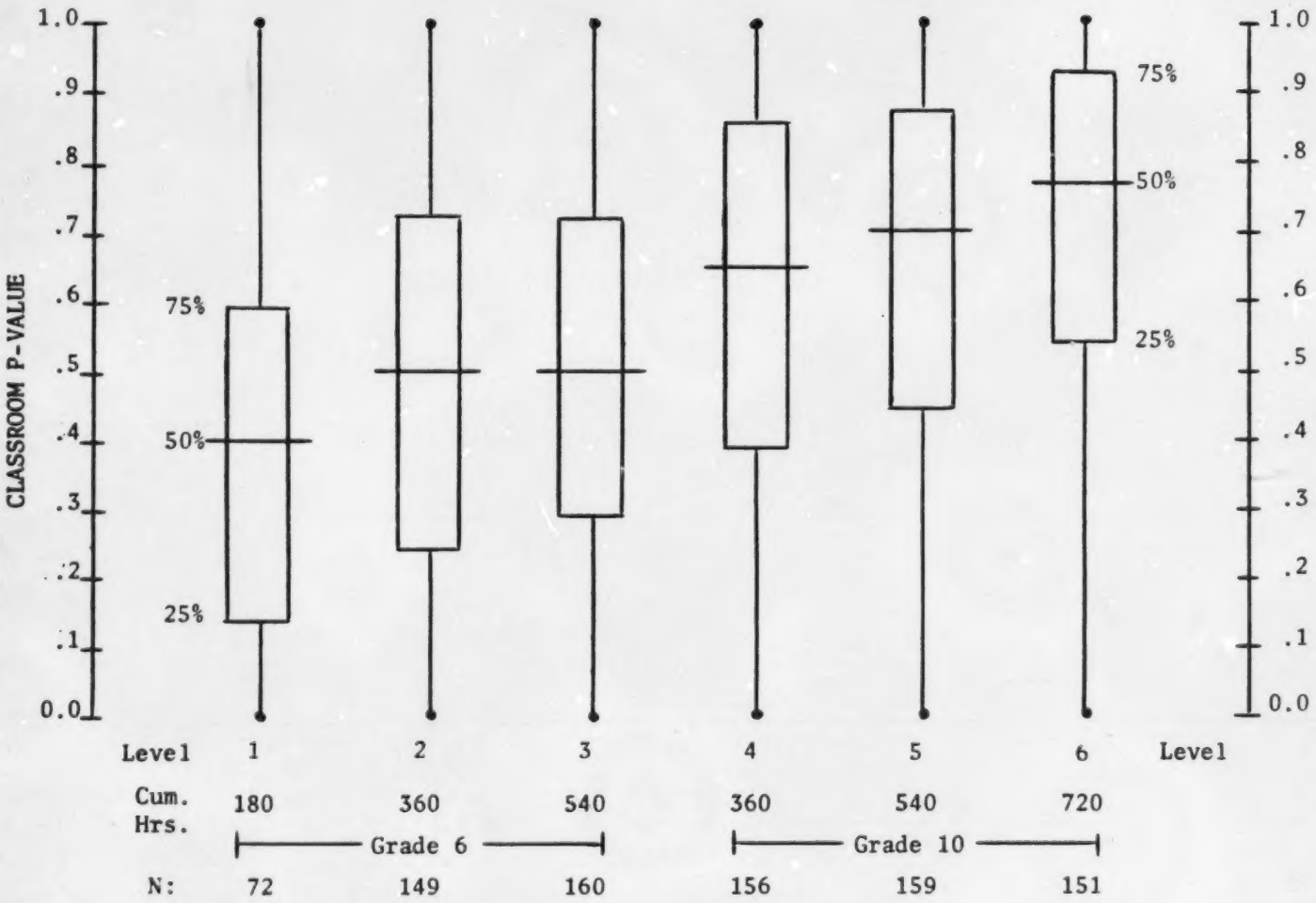
* Gr. 6 and Gr. 10 combined for significance testing.

Figure 1. Distribution of classroom p-values (proportion correct) by level for thirty-three French multiple-choice items.
N = no. of classrooms

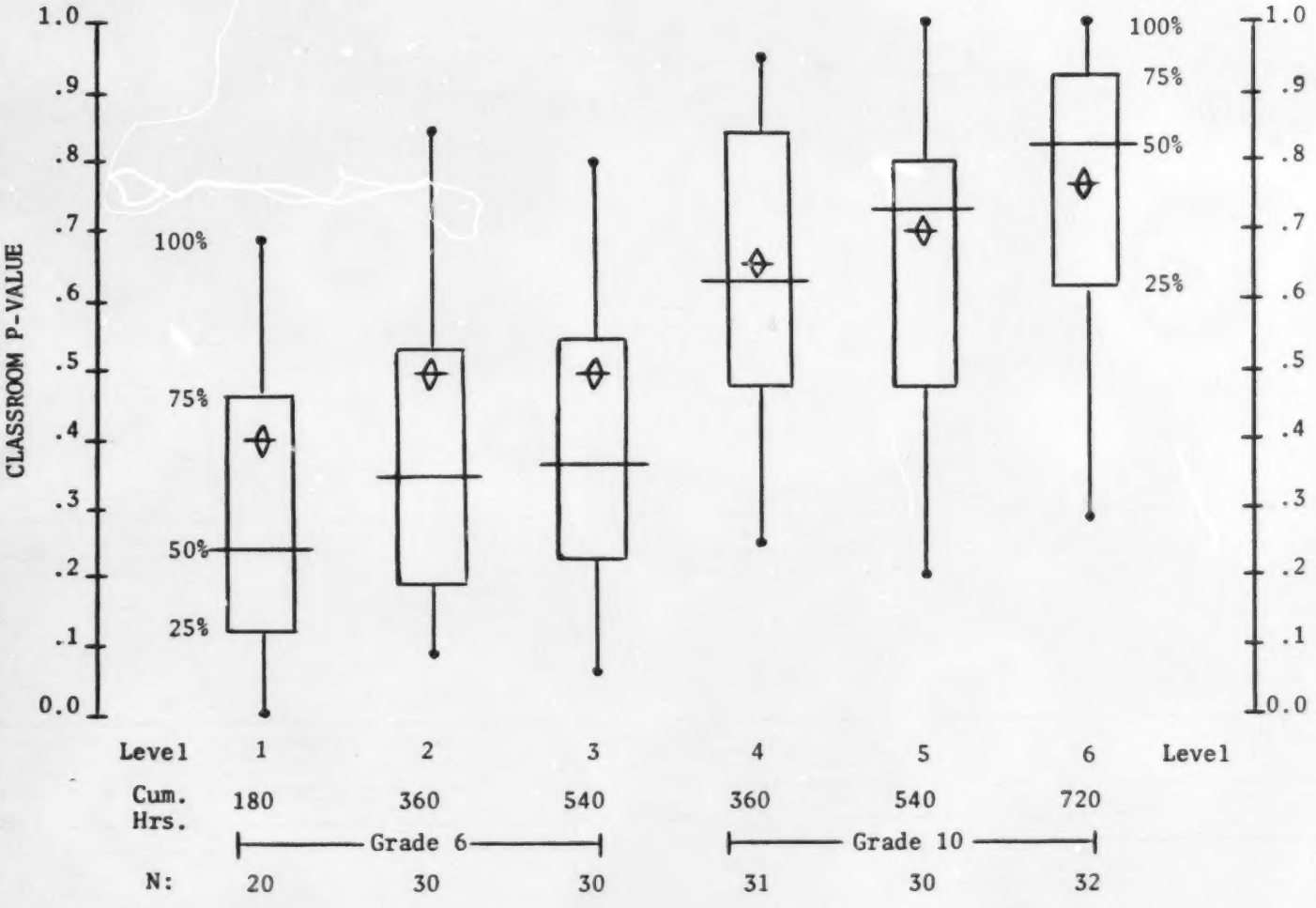
Figure 2. Distribution of classroom p-values (proportion correct) by level for seven French multiple-choice items where growth over levels was found.
⟠ indicates median of responses to all thirty-three such items (see figure 1).
N = no. of classrooms.

Figure 3. Gains in Speaking Skills. Cross-sectional results of contrasts among average ratings, item-by-item, for levels indicating grade and time.
*Number of items (speaking tasks). Some items were not scored on all dimensions. For some items, a different scale was used for grammaticality. These scores were not included.

Separating the effect of more hours . . .



Separating the effect of more hours . . .



Separating the effect of more hours . . .